Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

COMMENTARY

Imbalance *p* values for baseline covariates in randomized controlled trials: a last resort for the use of *p* values? A pro and contra debate

This article was published in the following Dove Press journal:
Clinical Epidemiology

Andreas Stang[1,2]
Christopher Baethge[3,4]

[1]Center of Clinical Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, Medical Faculty, University Hospital of Essen, Hufelandstr, Essen, Germany; [2]Department of Epidemiology, School of Public Health, Boston University, Boston, MA, USA; [3]Department of Psychiatry and Psychotherapy, University of Cologne Medical School, Cologne, Germany; [4]Editorial Offices, Deutsches Ärzteblatt and Deutsches Ärzteblatt International, Deutscher Ärzte-Verlag, Cologne, Germany

Background: Results of randomized controlled trials (RCTs) are usually accompanied by a table that compares covariates between the study groups at baseline. Sometimes, the investigators report *p* values for imbalanced covariates. The aim of this debate is to illustrate the pro and contra of the use of these *p* values in RCTs.

Pro: Low *p* values can be a sign of biased or fraudulent randomization and can be used as a warning sign. They can be considered as a screening tool with low positive-predictive value. Low *p* values should prompt us to ask for the reasons and for potential consequences, especially in combination with hints of methodological problems.

Contra: A fair randomization produces the expectation that the distribution of *p* values follows a flat distribution. It does not produce an expectation related to a single *p* value. The distribution of *p* values in RCTs can be influenced by the correlation among covariates, differential misclassification or differential mismeasurement of baseline covariates. Given only a small number of reported *p* values in the reports of RCTs, judging whether the realized *p* value distribution is, indeed, a flat distribution becomes difficult. If *p* values ≤0.005 or ≥0.995 were used as a sign of alarm, the false-positive rate would be 5.0% if randomization was done correctly, and five *p* values per RCT were reported.

Conclusion: Use of a low *p* value as a warning sign that randomization is potentially biased can be considered a vague heuristic. The authors of this debate are obviously more or less enthusiastic with this heuristic and differ in the consequences they propose.

Keywords: randomized controlled trial, distribution, statistical, random allocation

Introduction

Since its introduction into biomedical literature, null hypothesis significance testing (NHST) has caused much debate.[1–4] Despite many cautions, NHST remains one of the most prevalent statistical procedures in biomedical literature.[5,6] In 2016, Greenland et al reviewed overall 25 misinterpretations of NHST, *p* values, CIs, and power[7] and recently, the American Statistical Association released a policy statement on statistical significance and *p* values, including "The widespread use of "statistical significance" (generally interpreted as '*p*≤0.05') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process."[8]

Given the many warnings and misuses of NHST, it is unclear in which situations NHST can play a relevant role in the biomedical and epidemiologic literature. Here, we focus on the use of *p* values to assess imbalances of baseline covariates between

Correspondence: Andreas Stang
Center of Clinical Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Hufelandstr 55, 45147 Essen, Germany
Tel +49 201 9223 9290
Fax +49 201 9223 9333
Email andreas.stang@uk-essen.de

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/CLEP.S161508

study groups of a randomized controlled trial (RCT). In 1990, Greenland summarized the advantages of randomization as follows: 1) it makes estimates of effect "statistically unbiased, in that the statistical expectation (average) of the estimate over the possible results equals the true value" and 2) "it provides a known probability distribution for the possible results under a specified hypothesis about the treatment effect".[9]

Table 1 of results from an RCT usually presents baseline characteristics of included patients by treatment groups. These tables are sometimes accompanied by *p* values that are associated with the statistical null hypothesis of no baseline imbalances of covariates between the treatment groups (called "covariate imbalance *p* value", for the remainder abbreviated as CIP). If randomization was done properly, it can be expected that any baseline difference between treatment groups is solely due to chance. Epistemologically, it appears to be a paradox to test the null hypothesis of no imbalance if the mechanism that produced the covariate distributions of the treatment groups was a chance mechanism, that is, randomization. A valid randomization produces a flat distribution of the CIPs with *p* values ≤0.05 to be expected with a relative frequency of 5%. The aim of this debate is to illustrate the pro and contra of the use of CIPs for baseline covariates in RCTs.

Argument for the use of CIPs (Baethge)

For all its futility, criticism of NHST seems to be successful in one respect: it has become the norm not to report *p* values in table 1 of an RCT paper. A steady stream of literature discouraged NHST for baseline differences,[10–13] culminating in consolidated standards of reporting trials (CONSORT)'s elaboration document: "Unfortunately significance tests of baseline differences are still common […]. Tests of baseline differences are not necessarily wrong, just illogical. […]. Such hypothesis testing is superfluous and can mislead investigators and their readers."[14] So, CIP is bad practice and a sign that authors have no understanding of NHST and RCTs. Or have they?

The argument goes that if randomization went correctly, any distribution of variables among groups results from chance. But how can we be sure that randomization was correct? By a meticulous description of trial conduct, as CONSORT requires? Unfortunately, many authors do not follow: allocation concealment was adequately reported in merely 45% of all trials in journals endorsing CONSORT and in 22% of trials in journals not endorsing CONSORT.[15] Even if a trial looks good on paper, systematic error or fraud cannot be ruled out. Fanelli has meta-analytically estimated that 1 in 50 scientists self-reported to have "fabricated, falsified or modified data or results at least once".[16]

In fact, the literature provides many examples where low CIPs were a sign of scientific misconduct. Carlisle showed that in numerous RCTs by Yoshitaka Fujii – with 183 retractions, the frontrunner of Retraction Watch's "Leaderboard" – the CIPs were below 0.0001.[17] George et al found a *p* value of 1.9×10^{-17} regarding baseline weight in a "randomized" trial that was retracted later.[18] Even *p* values of almost 1 or exactly 1 can attract attention: they may be indicative of an improbable lack of variance. Kunz et al found *p* values of 0.997 or 0.988 too good to be true in the COOPERATE study, which was retracted in 2009.[19] In a historical case, Fisher calculated chi-square statistics from Gregor Mendel's publication in 1866, arrived at *p* values above 0.999, and concluded that Mendel had cheated[20] – a controversial claim. But it is undisputed that Mendel's results were biased.[20]

The *p* values alone cannot distinguish the reasons for baseline imbalances: chance or bias, including fraud. This, however, should not lead to discarding the *p* value as a warning sign. It is a screening tool with low positive-predictive value – the way fecal occult blood testing is a screening tool for colorectal cancer. Here is an example.[21] In a paper reporting an RCT on a modified cesarean section, the authors provided baseline characteristics of intervention and control groups that, when we recalculated the *p* values, were suggestive of bias, eg, for educational status. Under the assumption that randomization was correct in this trial, one would expect an imbalance between the two groups, as it was documented for educational status (or an even larger imbalance), with a probability of 0.00016. When we raised this point in a letter to the editor, the authors replied that parents were asked to participate not only before randomization but also after randomization and after they knew what treatment they were planned to receive. Further, at the same point in the study, staff was asked to participate.[22] This approach is different from the ethical imperative of allowing patients to withdraw their consent at any time. While initially the study used randomization, this approach introduces the strong possibility of allocation bias. In fact, it is a plausible explanation for baseline imbalance.

Low CIPs, therefore, should prompt us to ask for the reasons and for potential consequences: Can the way the trial was conducted explain the imbalance? Is the imbalance of prognostic importance and should it be adjusted for (advisable only when the extent of the imbalance is clinically relevant)? As often in medicine, there cannot be a hard-and-fast rule of when to dig deeper into CIPs, but very low CIPs (eg, below 0.005) and hints of methodological problems should certainly give pause for thought.

With regard to scientific misconduct, low CIPs as a screening tool should eventually give way to better means to detect fraud.[23–25] There are also more sophisticated methods of screening for fraud, eg, Carlisle's method,[17] but they are more difficult to apply. While statistically minded readers themselves can often calculate the *p* values from table 1, I fear this will not usually happen. Presenting CIPs may create a stronger incentive to discuss bias, or even worse, potential fraud. The way it is now, imbalances are often not discussed (eg, Schramm et al[26]). In contrast to other applications of NHST and under the assumption of sufficient study size, CIPs are what we need in evaluating randomization: the probability of the observed or a stronger baseline covariate imbalance if chance was the only explanation. It seems odd that the *p* values have become outlawed in precisely one of the few places where they should have a role.

Arguments against the use of CIPs in RCTs (Stang)

A fair randomization produces the expectation that the distribution (sic) of CIPs follows a flat distribution. It does not produce an expectation related to a single CIP. In contrast, Baethge does not use the distributional expectation, but uses an expectation related to single CIPs.

Furthermore, the distribution of CIPs in RCTs can be influenced by three factors, so that the expectation of a flat distribution of CIPs is not met anymore. First, the CIP distribution becomes distorted if the baseline covariates for which the *p* values are calculated are correlated with each other. For example, Flaherty et al presented the baseline characteristics of 322 randomized patients with metastatic melanoma with a *BRAF* mutation who either received trametinib or chemotherapy. They presented the percentage of "disease at ≥3 sites" and the percentage of "history of brain metastasis" as 57% versus 52% and 4% versus 2%, respectively, for the two treatment groups without *p* values. These two baseline characteristics are associated with each other. Patients with disease at ≥3 sites have a higher probability to have a history of brain metastasis than patients with disease at <3 sites.[27] Second, unblinded study teams of RCTs can produce differential misclassification or differential mismeasurement of baseline covariates. This differential bias also influences the distribution of CIPs. Third, the median number of CIPs presented in the tables of published RCTs is 16, which makes the study of the CIP distribution quite unreliable.[12] What does the reader learn about the distribution of CIPs if table 1 of the published RCT contains only a few CIPs with one out of them being below 0.005? For judging whether the realized CIP distribution in an RCT is, indeed, a flat distribution, the presentation of only a few CIPs in table 1 of an RCT is not sufficient evidence for this judgment. At best, investigators would present as many as possible CIPs graphically illustrated in a supplementary figure.

In addition, it is unclear to me what Baethge's approach implies for CIPs between 0.005 and 0.995. Can they be considered as an all-clear signal?

A brief review of the RCTs published in the *New England Journal of Medicine*, *Lancet*, and *JAMA* (Medline search: randomized controlled trial [publication type] AND ["JAMA" {journal} OR "N Engl J Med" {journal} OR "Lancet" {journal}] AND 2017/05:2017/06 [dp]) for the months May and June 2017 revealed overall 57 published RCTs. With the exception of one RCT, all RCTs refrained from providing CIPs in table 1. Overall, 27 (47%) of the RCT papers provided a statement about the presence of any statistically significant imbalance and reported only those CIPs that were significant at α=0.05. Eight out of these 27 RCT papers found statistically significant differences for at least one baseline covariate. Another 18 RCT papers (32%) only gave qualitative statements about imbalances at baseline. Interestingly, 19% did not provide any statement about baseline imbalances. Only one paper actually reported CIPs for all baseline covariates presented in table 1. This mini review shows that CIPs are only rarely presented nowadays in RCT publications of top medical journals. Obviously, for the use of the proportion of CIPs being ≤0.005 as a quality control measure of the randomization, a substantial number of *p* values should be presented to learn anything about the CIP distribution. If 16 *p* values are published per RCT[12] and a *p* value of ≤0.005 is interpreted as a warning, then the false-positive rate is 8% for studies where randomization has been properly performed. This rate increases to 16% if one interprets the *p* values which are ≥0.995 as a warning.

Conclusion

The study of CIPs in RCTs to detect potential bias related to the random assignment of the treatments may be called a heuristic ("rule of thumb"). According to the Cambridge Dictionary of Philosophy, a heuristic is defined as "A rule adopted to reduce the complexity of tasks; a heuristic may not reach a solution even if there is one, or may provide an incorrect answer (as opposed to an algorithm, ie, mental short cut)."[28] The study of the distribution of CIPs in RCTs does not reach a solution, or it even may provide an incorrect answer in case of correlated baseline covariates or differential misclassification or differential mismeasurement of baseline covariates. Besides these theoretical objections, the heuristic is hampered by the fact that tables of RCTs usually contain only a few covariates, so that a distribution of CIPs is hard to study. One is left with a kind of "cherry picking" of very low CIPs when they are reported. The authors of this debate are obviously more (Baethge) or less enthusiastic (Stang), with the former advocating the presentation of CIPs, its careful use as a screening tool, and its interpretation within the context of each study, while the latter emphasizes the dangers of misuse. The aim of this debate was to further trigger the discussion of the role of NHST in biomedical research that uses randomization.

Statistical theory teaches us that randomization produces balance of baseline covariates in the long run, that is, over an infinite series of RCTs, but not necessarily in a single RCT. Therefore, a baseline imbalance of a prognostic factor in a single RCT due to chance is not a sign of bias. However, if chance produces imbalance of covariates, investigators consequently adjust for baseline imbalances, as imbalances by chance also produce mixing of effects.[29]

Our debate is centered on the appropriateness of p values as a screening tool for imbalanced baseline covariates. It is noteworthy that other more elaborate approaches have been proposed for the investigation of[23–25] and the adjustment for baseline imbalances, for example, propensity scores (PS). The individual PS refers to the probability for a subject in the study of being assigned to the intervention arm A rather than intervention arm B, given the patient's characteristics at baseline. Leyrat et al proposed a c-statistic of the PS model to detect global baseline covariate imbalance in cluster RCTs. In the absence of baseline imbalance, the c-statistic of the PS model is expected to be close to 0.5. In the presence of imbalance, this c-statistic will be larger than 0.5. This procedure is still being tested and there remain unresolved questions in dealing with this procedure. For example, it is not clear how large the c-statistic has to be to decide that a relevant baseline imbalance is present.[30]

For the detection and judgment of imbalances between the study groups, it remains important that descriptive statistics of the groups (categorical characteristics: percentage values; continuous characteristics: eg, mean values and SDs) are presented. Whether a baseline imbalance is meaningful or not depends on subject matter knowledge. For example, it is clinically relevant in a stroke prevention study if 30% diabetics are in one arm of the study and only 15% are diabetics in the other arm, regardless of the p value, as diabetes mellitus is a very relevant risk factor for stroke.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Boring EG. Mathematical vs. scientific significance. *Psychol Bull*. 1919;15:335–338.
2. Hogben LT. *Statistical Theory: An Examination of the Contemporary Crisis in Statistical Theory From a Behaviourist Viewpoint*. London: George Allen & Unwin; 1957.
3. Morrison DE, Henkel RE. *The Significance Test Controversy: A Reader*. Chicago, IL, USA: Aldine Pub; 1970.
4. Cohen J. The earth is round (p<0.05). *Am Psychol*. 1994;49(12): 997–1003.
5. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting *P*-values in the Biomedical Literature, 1990–2015. *JAMA*. 2016;315(11): 1141–1148.
6. Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstracts of major medical and epidemiology journals 1975–2014: a systematic review. *Eur J Epidemiol*. 2017;32(1):21–29.
7. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P*-values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350.
8. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Statistician*. 2016;70(2):129–133.
9. Greenland S. Randomization, statistics, and causal inference. *Epidemiology*. 1990;1(6):421–429.
10. Altman DG, Dore CJ. Baseline comparisons in clinical trials. *Lancet*. 1990;335(8682):149–153.
11. Senn S. Testing for baseline balance in clinical trials. *Stat Med*. 1994;13(17):1715–1726.
12. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064–1069.
13. Knol MJ, Groenwold RH, Grobbee DE. *P*-values in baseline tables of randomised controlled trials are inappropriate but still common in high impact journals. *Eur J Prev Cardiol*. 2012;19(2):231–232.
14. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869.
15. Turner L, Shamseer L, Altman DG, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev*. 2012;11:MR000030.
16. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*. 2009;4(5):e5738.
17. Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia*. 2012;67(5):521–537.
18. George BJ, Brown AW, Allison DB. Errors in statistical analysis and questionable randomization lead to unreliable conclusions. *J Paramed Sci*. 2015;6(3):153–154.
19. Kunz R, Wolbers M, Glass T, Mann JF. The COOPERATE trial: a letter of concern. *Lancet*. 2008;371(9624):1575–1576.
20. Pires AM, Branco JA. A statistical model to explain the Mendel-Fisher controversy. *Stat Sci*. 2010;25(4):545–565.
21. Baethge C, Blettner M, Friese K. Armbrust et al. 2015: randomization questionable. *J Matern Fetal Neonatal Med*. 2016;29(22): 3730–3731.
22. Armbrust R, Henrich W. Re: the Charite cesarean birth: a family-orientated approach of cesarean section. *J Matern Fetal Neonatal Med*. 2017;30(1):43–45.
23. Buyse M, George SL, Evans S, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med*. 1999;18(24):3435–3451.

24. Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ*. 2005;331(7511):267–270.
25. van den Bor RM, Vaessen PWJ, Oosterman BJ, Zuithoff NPA, Grobbee DE, Roes KCB. A computationally simple central monitoring procedure, effectively applied to empirical trial data with known fraud. *J Clin Epidemiol.* 2017;87:59–69.
26. Schramm E, Kriston L, Zobel I, et al. Effect of disorder-specific vs nonspecific psychotherapy for chronic depression: a randomized clinical trial. *JAMA Psychiatry*. 2017;74(3):233–242.
27. Flaherty KT, Robert C, Hersey P, et al; METRIC Study Group. Improved survival with MEK inhibition in BRAF-mutated melanoma. *N Engl J Med*. 2012;367(2):107–114.
28. Audi R. *The Cambridge Dictionary of Philosophy*. 2nd ed. Cambridge: Cambridge University Press; 1999.
29. Rothman KJ. Epidemiologic methods in clinical trials. *Cancer*. 1977;39(4 Suppl):1771–1775.
30. Leyrat C, Caille A, Foucher Y, Giraudeau B. Propensity score to detect baseline imbalance in cluster randomized trials: the role of the c-statistic. *BMC Med Res Methodol*. 2016;16:9.